# On the sample size of clinical trials

Papageorgiou, Spyridon N

DOI: https://doi.org/10.1080/14653125.2018.1501929

**STATISTICAL CORNER**

# On the sample size of clinical trials

Spyridon N. Papageorgiou

Clinic of Orthodontics and Pediatric Dentistry, Center of Dental Medicine, University of Zurich, Zurich, Switzerland

**ORCID** Spyridon N. Papageorgiou http://orcid.org/0000-0003-1968-3326

**CONTACT** Spyridon N. Papageorgiou Clinic of Orthodontics and Pediatric Dentistry, Center of Dental Medicine, University of Zurich, Plattenstrasse 11, Zurich CH 8032, Switzerland; snpapage@gmail.com.

Words in text: 959

## Disclosure statement

No potential conflict of interest was reported by the author.

**MANUSCRIPT**

**Theoretical scenario**

Clinical trials in orthodontics often have as primary outcome the total duration of fixed appliance treatment from bonding to debonding of the appliances. This is one of the most clinically relevant outcomes to the patient or the orthodontist and prolonged treatment times are associated with adverse effects like patient burnout and the development of white spot lesions or root resorption. Data from existing randomized trials indicate that the outcome of treatment duration is normally distributed and that an average duration of 20 months with a Standard Deviation (SD) of about 5 months might not be an unrealistic estimate to expect—although large variability in treatment times exists.

A group of orthodontic researchers plan to conduct a two-group parallel randomized clinical trial to assess the effect of corticotomy on the reduction of overall treatment duration. However, disheartened from the large sample size needs yielded by a formal statistical sample size calculation, they arbitrarily decide to recruit a total of 20 patients in the two treatment groups and try to pass the study as a 'pilot study'. Their justification is that as this is a randomized trial the possibility of bias will be low and their results will still be robust, despite the objectively small sample size.

After completing their trial and collecting data, they routinely check for normal distribution of treatment duration with the Shapiro-Wilk test, which gives them a P value of 0.01 that indicates absence of normality. Therefore, they adjust their descriptive and inferential statistics to the non-normal distribution. Finally, they conclude that although this is a pilot study with small sample, the randomized nature of the study means that its results are robust and in low risk of bias.

**Which of the following statements are correct, if any:**

(a) Having a small sample in a trial cannot affect the average treatment duration measured.

(b) Having a small sample in a trial might affect the precision of the estimated results.

(c) Having a small sample in a trial cannot affect the 'normality' (or lack thereof) of the trial's outcome.

(d) Randomization can safeguard against any limitations associated with a small sample size.

**Discussion**

To illustrate the first statement, we generate a large parent sample of 200 patients that should correspond to the 'true' mean duration of 20 months and an SD of 5 months. As all samples are subject to a certain variability, we end up with the sample of 200 patients having a mean of 19.4 months and an SD of 5.4 months, which is pretty close to what we initially aimed. If we now draw random trial samples of 20 patients each from the parent sample, we can see an interesting finding in **Figure 1**. Although many of the drawn samples (blue boxes) are relative close to the 'true' average duration of 20 months (green box), the mean values of the trials vary considerably—like sample H that has a mean duration of 18.2 months. This is exaggerated even more if we further reduce the recruited sample size from 20 to 10 patients, where the mean values of the small trials (red boxes) are generally more far away from the true mean (green box). We see therefore signs of *bias of the estimator*, in which the sample size of a trial together with pure chance can influence the results that are actually measured in a trial. Therefore, statement A is false.

Furthermore, we can assess the precision of the measured results in the drawn samples by calculating the 95% Confidence Intervals (CIs) of the mean treatment duration in each trial. We would then see that the parent sample A has a very narrow 95% CI spanning only about 1.5 month (19.2 to 20.7 months), which indicates high precision (Figure 1). On the other side, trials with 20 patients (coloured blue) have much wider 95% CIs of at least 4 months. This is again further exaggerated in trials with 10 patients (coloured red), where the 95% CIs span almost 10

months! We see therefore, that even in a perfectly random setting, the sample size of a trial is closely related to the trial's precision—and statement B is correct.

In order to assess the truth of statement C, one method would be to plot the distributional histograms of the drawn trial samples and perform the Shapiro-Wilk test to formally assess the normal distribution of treatment duration. As **Figure 2** indicates, the distribution of the parent sample A resembles that of a normal distribution (top) and this is confirmed by a non-significant Shapiro-Wilk test ($P>0.05$). On the other side, a great variability in the outcome distribution can be seen among both trials with 20 patients (middle row) and with 10 patients (bottom row). Also, formal testing for normality with the Shapiro-Wilk test gives diverse results and indicates that some trials from the middle or bottom row have non-normally distributed outcome ($P<0.05$). This would lead the authors of the trial to conclude that treatment duration is not normally distributed and chose a different statistical analysis plan than better fits normally-distributed data. Therefore, statement C is wrong.

Finally, randomization in a trial is employed to ensure the baseline equivalence of the randomized groups for all known or unknown factors that could potentially affect treatment duration. This means that any difference in the final treatment duration of the randomized groups could be potentially attributed to their different treatment protocol. Randomization cannot safeguard from the limitations of having a small sample, which can ultimately affect the results of the trial and their validity. This is the reason why the CONSORT (Consolidated Standards of Reporting Trials) statement (Moher et al. 2010) has distinct items for describing sample size calculation and randomization, as these are two separate procedures independent of each other.

## References

Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG; Consolidated Standards of Reporting Trials Group. 2010. CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol. 63:e1–37.

**Figure 1.** Results of the same hypothetical trial with a set mean treatment duration of 20 months and a standard deviation of 5 months. The trial's sample size varies between 200, 20, and 10 patients and new samples are randomly drawn. The results of each trial are expressed as the mean with its 95% Confidence Interval (CI).

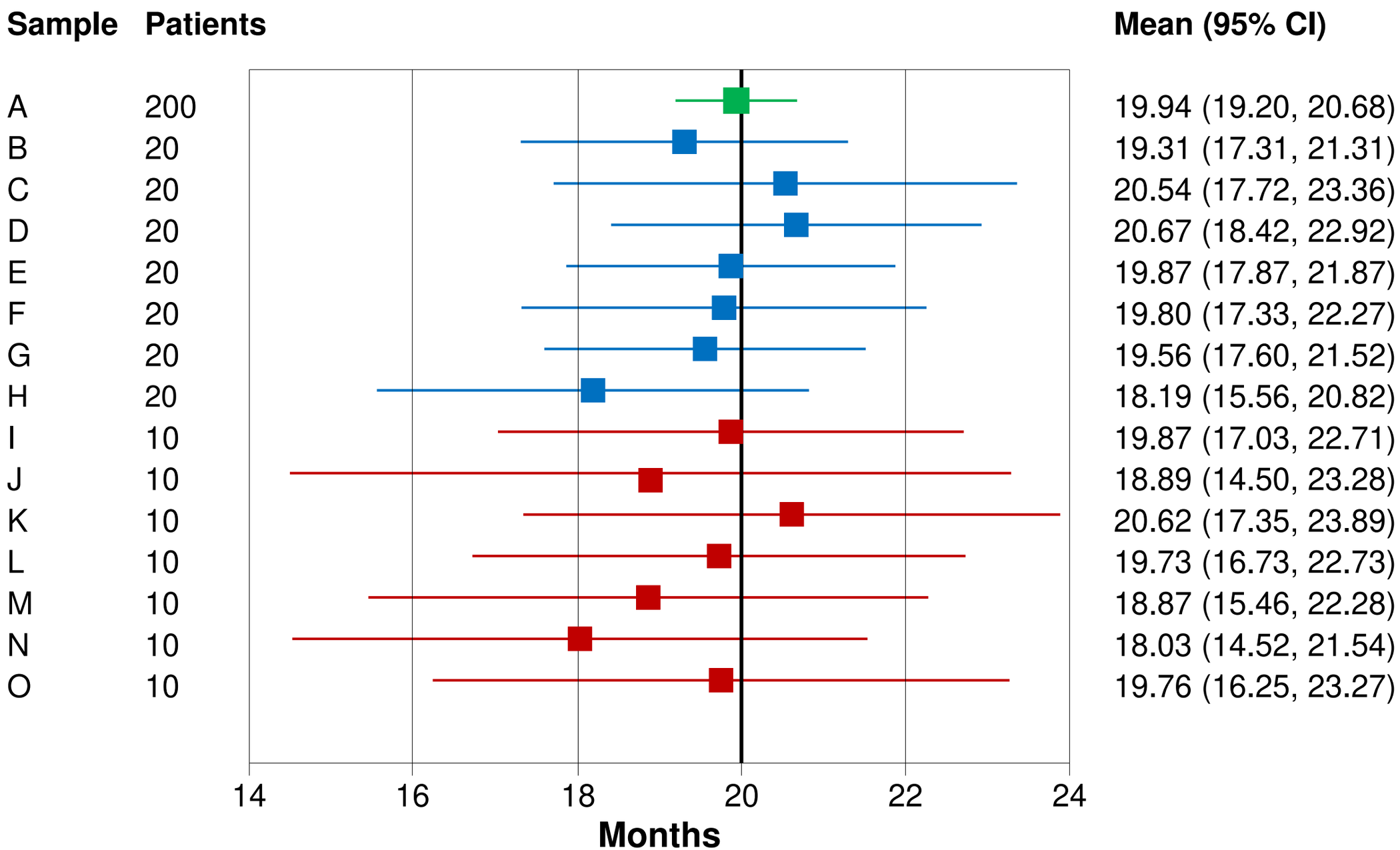

**Figure 2.** Distributional histograms of the samples from Figure 1. The normal distribution in each case is checked with a P value from the Shapiro-Wilk test.